学术论文创新贡献句识别研究*

罗卓然 蔡乐 钱佳佳 陆伟
武汉大学信息管理学院 武汉 430072

摘 要:[目的/意义]学术论文贡献句是体现论文创新性和学术价值的重要形式。以学术论文全文本和MeSH主题词为数据基础,利用自然语言处理和深度学习技术,实现学术论文贡献句识别,为学术文本创新贡献内容的细粒度挖掘奠定基础,对实现基于认知计算的学术论文评价具有重要的理论和现实意义。[方法/过程]首先,以PubMed论文全文本为数据来源,抽取论文Mesh主题词,对论文贡献句进行要素分析和特征提取。其次,采用半自动方式实现标注数据。最后,基于Albert深度学习模型实现贡献句的自动识别。[结果/结论]通过数据一致性检验证明实验标注的训练数据的可信性,实验结果表明,相较于其他深度学习模型,训练的自动识别模型能够更有效识别学术论文中贡献句。

关键词:贡献句 学术论文 创新性 Albert
分类号:G251.2
DOI:10.13266/j.issn.0252-3116.2021.12.009

1 引言

近年来,科技评价的改革日益受到有关部门和学术界的关注。2020年9月,习近平总书记在科学家座谈会上强调要依靠改革激发科技创新活力,通过深化科技体制改革把巨大创新潜能有效释放出来,坚决破除"唯论文、唯职称、唯学历、唯奖项"[1]。对传统论文评价方式的革新将有利于正确引导科研方向,并进一步释放我国科研活力和创新潜力。科技评价是对创新成果的学术贡献、学术价值、学术影响以及社会影响、应用价值等的综合评估。

学术论文的贡献是论文创新性、科学性和学术价值的集中体现,学术论文的贡献价值体现在作者提出的新理论、新方法、新技术、新成果、新应用等创新贡献要素对人类社会发展与科技进步所带来的社会价值与经济效益。学术论文创新性评价是学术评价的一项重要任务,创新的评价与度量是一项复杂的系统工程,其中包括创新性本身的复杂性,以及评价的过程和要素的多样性和不确定性。创新评价与度量旨在评价创新的"意义"或"价值",即该研究对已有的科研环境和知识体系所作的研究贡献。可见学术论文的创新是论文贡献价值的核心要义,创新是具有研究贡献的论文中不可或缺的重要组成部分。由于衡量论文的创新性需要综合考虑多种因素,创新的发现存在一定的时滞性和不确定性,导致科研创新发现工作往往具有严重的滞后性。但是,在创新内容的发现过程中,创新要素的描述通常体现了论文的贡献价值,对单篇学术论文而言,若要实现内容层面的创新性评价,需要准确地找到学术论文中的贡献描述句,即直接描述或能体现潜在创新价值的句子。

目前,国内外关于学术论文创新贡献点抽取或识别的直接研究较少,相关研究主要体现在创新点识别、亮点句识别、创新研究评价句、方法句识别等方面。贡献句是论文创新内容的重要承载形式,有效地识别论文的学术贡献是创新评价研究的前提,将为创新内容的精准识别提供理论与数据基础。医学主题词表(Medical Subject Headings,MeSH),是美国国立医学图书馆编制的权威性主题词表,其提供主题词的自然信息(词义、同/近义词、可组配副主题词等),能够准确揭示文献内容的主题,与论文贡献描述内容密切相关。本研究以学术论文的全文本和MeSH主题词两类关键信息为数据对象,基于自然语言处理和深度学习技术,

* 本文系国家社会科学基金重大项目"基于认知计算的学术论文评价理论与方法研究"(项目编号:17ZDA292)研究成果之一。
作者简介:罗卓然(ORCID:0000-0003-0677-8350),博士研究生,E-mail:zoraluo@whu.edu.cn;蔡乐(ORCID:0000-0003-1278-4343),硕士研究生;钱佳佳(ORCID:0000-0002-6058-1287),硕士研究生;陆伟(ORCID:0000-0002-0929-7416),教授,博士生导师。
收稿日期:2020-12-10 修回日期:2021-03-25 本文起止页码:93-100 本文责任编辑:徐健

开展了基于深度学习预训练模型的学术论文贡献句识别研究。为实现论文创新贡献句的抽取，本研究主要开展了两方面的工作，一是从贡献主题词的角度出发，利用自然语言处理技术提取论文贡献句语法特征，提出了论文贡献句的抽取方法；二是基于深度学习技术训练 Albert 论文贡献句分类模型，实现了对论文贡献句的分类与识别。

2 相关研究

2.1 学术论文创新贡献句识别

学术论文是具有新的科学研究成果或创新见解和知识的科学记录，某种已知原理应用于实际中而取得新进展的科学总结，用以提供学术交流与讨论的材料，或是发表在学术刊物上，或作其他用途的书面文件[2]。学术论文的重要特征之一是论文创新贡献价值，即论文中的观点、理论、方法等内容要素是否具有发现新的问题、解决现有难题、促进学科发展等方面的贡献价值。

目前，国内外关于学术论文创新贡献点的直接研究较少。李如森等认为科技论文的创新点分布在文章的主题、技术背景、技术方法、结论等部分[3]，体现出科技论文的创新点的分布并不限于特定的章节部分，而是可能出现在论文全文的各个部分。温有奎等[4]提出论文创新点动态挖掘模板，以句子中的特征词作为抽取特征项，实现科技文献中科研创新点碎片的动态挖掘。张帆等[5]以领域词表和本体中的关系为基础，实现了基于识别规则和补充规则对论文中创新句的抽取。索传军等[6]研究了学术论文中描述核心观点的亮点句的特征和规律，将亮点分为研究创新型亮点、研究方法型亮点、研究过程型亮点与研究结论型亮点4类，得出亮点句主要分布于论文的研究结果与研究方法部分，并在各个章节中无序随机分布。章成志等[7]以图书情报档案学科为例，通过基于规则的方法抽取了创新研究评价句，将评价句分为概念理论类、观点发现类、模型方法类、派别领域类、系统软件类和实践应用类6种类型，发现评价句主要与概念理论相关且较多处于论文靠前的位置。曹树金等[8]从句子级创新性识别出发，将句子的创新类型总结为理论创新、观点\概念创新、研究方法创新、研究问题\对象创新4大类，抽取了国内外两种期刊的论文的创新表达范式。温浩[9]首先根据句法和语义功能利用支持向量机将科技论文的摘要分为6类，然后对不同类别的数量分布和句子位置进行统计，并分析了句子类型和句子语义位置结构特征。周海晨等[10]结合深度学习和规则的方法，提出了学术创新贡献识别方法以识别文章中的创新短语，但仅在少量数据集上进行了模型训练，并未仔细阐述创新短语和贡献短语这两个类别标签之间的差异。L. L. Chen 等[11]利用词性标记的方法提取 N-gram 作为候选单词，并通过检查 Scopus® 数据库以确定其是否出现过从而判断主题词的创新贡献价值。J. Allan 等[12]认为新词很可能揭示论文作者所提出的新概念、新指标以及做出的新贡献等，利用句子中出现新词的个数，筛选文本中的新颖性句子，在 TREC 2002 新颖探测任务中取得了较好的效果。S. Teufel[13]等利用学术论文写作中的修辞现象，通过文本提取的方法，抽取或总结论文对研究背景的创新贡献，但是存在较多的噪声单元，分类的准确性相对较低。K. Heffernan[14]等将学术研究的贡献定义为学术文本中的问题及对应的解决方案，利用机器学习的方法定义了一组与目标类别相关的15个特征，在 ACL 数据集种可以较好地区分问题、非问题、解决方案。

2.2 学术文本抽取与表征

学术文本贡献句识别研究主要用到文本信息抽取与识别技术。常见的信息抽取主要包括两方面内容，即目标属性的抽取和目标之间关系的抽取。信息抽取领域的国际评测会议 Message Understanding Conference 制定了具体的任务和严格的信息抽取评估体系，核心内容包括命名实体识别、共指消解、关系抽取、事件抽取等具体内容。目前，学术文本信息提取最主要的方法包括基于知识的方法和基于机器学习的方法。基于知识的方法是依靠领域专家编制规则，将相应实体加入预先编制好的框架中，使系统能处理特定的信息抽取问题。例如冷伏海等[15]首先阅读高质量领域综述性文献，对科技文献进行语义标注，得到领域相关学术术语，制定相应规则抽取文献中领域研究相关的关键性能指标。毛琛瑜等[16]通过句式分析、词频统计、共现分析等方法，分析中文科技文献中新发现语言描述模式，找到了新发现语言的特征搭配。

文本抽取是文本表示的前提，在抽取出特征文本后，需要对非结构化的文本进行字词编码，将其转换为计算机可识别、可计算的数值形式，即对文本进行向量化表示。最早使用的文本向量化表示方法是独热（One-Hot）编码，该方法将文本划分为独立的单词，在词汇表中每个单词被表示为索引位置为1，其他位置为0的向量。该方法的特点是简单，但没有考虑单词之间的联系和相似性，不包含单词之间的语义相似性。

针对此，T. Mikolov 提出了分布式表示模型 Word2vec[17]，对词汇与词汇之间的语义关系进行建模，与 One-Hot 向量不同，词向量是一个维度较低的稠密向量，词向量与神经网络的结合，大大促进了自然语言处理任务的效率和效果。词向量分为静态词向量和动态词向量，静态词向量在上下文发生变化时也只能表示一个单词，例如 Word2Vec 和 Glove[18] 模型，无法解决一词多义的问题；而动态词向量会根据词的上下文动态地调整词向量。动态词向量包括一些预训练模型，如 ElMo[19]（Embedding from Language Models）、BERT[20]（Bidirectional Encoder Representations from Transformers）等。预训练语言模型能够基于上下文捕获词语的深层语义信息，通过大规模语料训练学习到的特征对词语进行上下文特征表示。鲁威[21] 对多因素的文本分类进行了研究，利用 Elmo 模型根据上下文语境的不同将词映射为不同的向量，验证了 Elmo 动态词向量相较于静态词向量的优势。顾亦然等[22] 针对电击领域专业实体识别困难、精度低等问题，利用 BERT 模型捕获上下文语义表示动态生成词向量，并结合双向长短记忆神经网络（BiLSTM）和条件随机场（CRF）实现了中文命名实体识别，实验证明该方法优于其他算法模型，能有效解决该领域实体边界模糊且难于识别的问题。廖胜兰等[23] 基于对话系统中的意图分类问题，采用预训练模型和知识蒸馏等技术，提出了一个基于 BERT 模型的知识蒸馏意图分类模型，在原有数据和计算资源的基础上将意图分类的准确率提升 3.8%。

2.3 Albert 预训练语言模型

在 BERT 出现之前，预训练模型多为单向模型，如 GPT 单向训练模型，ELMo 模型虽然是双向但训练过程是分开的。BERT 是一个完全的双向语言模型，其训练结果表明双向语言模型相较于单向语言模型对文本语义的理解更加深刻。BERT 是一种基于 Transformers 结构的双向语言模型，在预训练任务中采用了掩码语言模型（MLM，Masked Language Model）和下一句预测（NSP，Next Sentence Predict）。

2019 年，谷歌的 Z. Z. Lan 等发现当 BERT 模型复杂到一定的程度时，随着模型参数增加，模型的训练效果反而会下降，为此其提出了 Albert（A Lite BERT）[24] 模型，该模型在 BERT 模型的基础上做了模型压缩与优化，使其能够在参数规模上得到降低，同时在一定程度上提升模型训练效果。Albert 在 BERT 的基础上引入了 3 种优化策略：因式分解嵌入层矩阵（Factorized Embedding Parameterization）、跨层参数共享（Cross-layer Parameter Sharing）和句子顺序预测（Sentence-order Prediction，SOP）。上述改变使得 Albert 成为自然语言处理任务中效果最出色的模型之一，在数据量较小的情况下该模型的优势也更加显著。本研究通过使用 Albert 模型完成贡献句的句子特征抽取，训练学术论文贡献句识别模型，达到识别论文中贡献句的目的。

3 学术论文贡献句内涵与筛选

3.1 学术论文贡献句内涵

学术论文的贡献（contribution）是指当前的研究能对现有理论、实践作出的改进、完善与应用，通常贡献点包含论文的研究意义、研究价值、研究影响等方面，是论文价值的集中体现。论文贡献内容是论文中存在与现有文献不同的知识单元文字描述、公式算法和图像表格等论文元素，反映在创新点的创新价值与贡献程度两个方面。

在创新贡献的要素与评价研究方面，C. J. Denholm[25] 从单个学科的维度对创新贡献进行了界定，并从不同学科的视角总结了博士论文创新评价指标，指出不同学科之间由于学科自身特质、学科文化等差异，对于创新性标准的理解和评价存在较大差异。T. Dahl[26] 构建了一个表征论文新的研究贡献的特征词表，以自动识别论文中的创新贡献点，然而该方法依赖领域专家且不能涵盖所有的特征和规则。李瑛[27] 等认为科技期刊论文创新贡献的合理呈现必须包括 8 个要素，即创新方面、创新类型、创新内容、创新程度、创新质量、创新价值、创新缺陷和表达位置。李贺等[28] 以知识元研究为基础，从内容角度将学术论文创新分为研究问题创新、理论创新、方法创新及结论创新 4 个维度，基于其构建了学术论文知识元本体模型和创新性评价方法，并对《图书情报工作》2015 年至 2017 年发表的文章的创新性进行度量，但存在部分论文的测度结果创新性得分为 0 的情况，计算方法上还有待调整，但该研究提出了一套丰富的理论模型，在理论方法层面为本研究的贡献句智能识别研究提供了一定的参考。

作为科研工作者创新成果的载体和文字表述，学术论文创新贡献句是知识创新贡献、技术创新贡献、应用创新的文字体现，具备创新贡献价值的成果应具有科学性、新颖性、实用性等特点。目前对学术论文创新贡献句的内涵还没有统一的界定，但通过文献梳理和调研发现学术论文中的贡献句的分布与描述具有以下

特点：

（1）从贡献内容来看，即论文所承载的具体研究贡献，学术内容贡献句内容维度如图 1 所示，具体内容包括：①研究问题贡献，即开辟新的研究领域，提出新的论题或研究对象，发现开创性的研究问题，弥补研究空白或瓶颈的研究突破，或对现有研究不足与存在问题的改进与完善；②理论研究贡献，即针对已有的问题提出新的观点、见解、思路、理论模型或框架，发现新规律、提出新假说或新理念；③研究方法贡献，提出了研究问题或解决问题的新视角，采用了新的试（实）验和操作方法、论证或计算方法，引进或提出新的技术或方法；④成果与应用贡献，做出了新的发明或设计，现有的方法或技术应用在新领域中，或者拓展了其应用范畴。

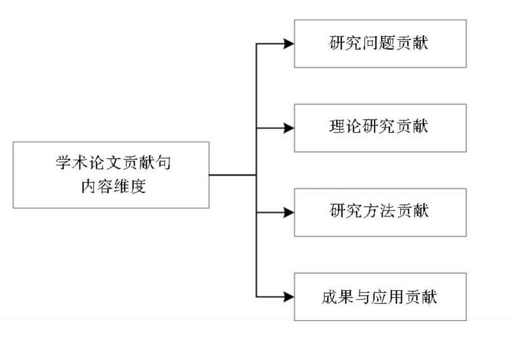


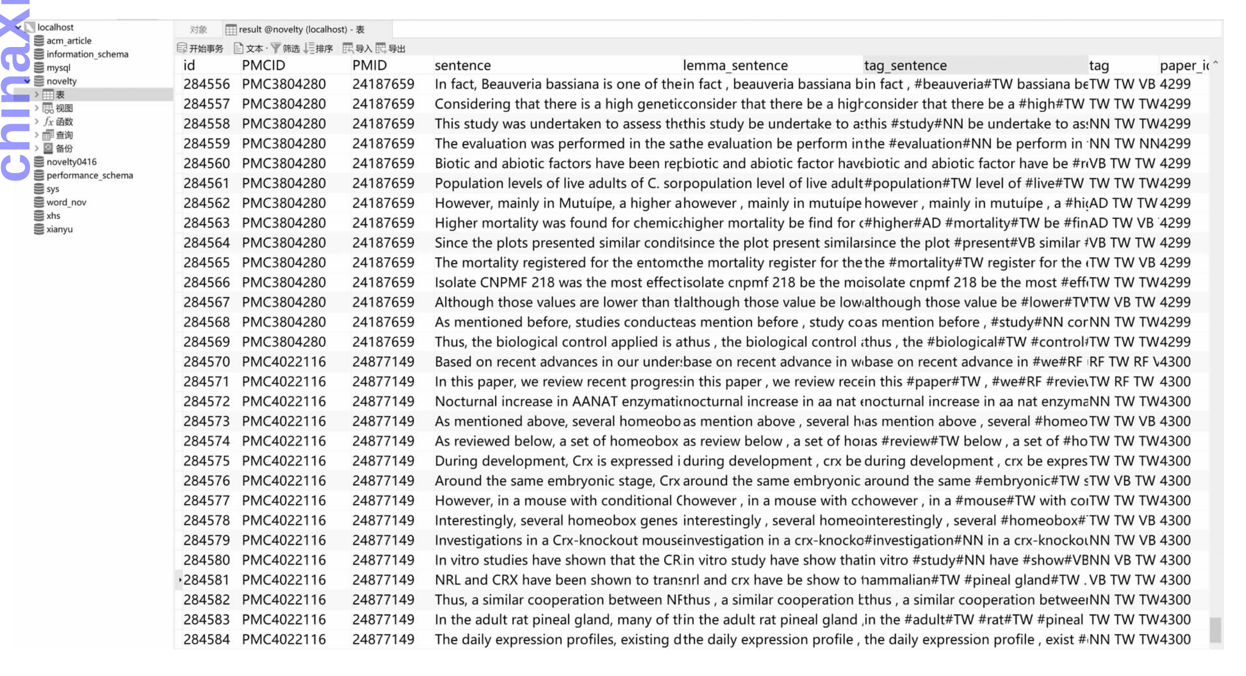
图 1　学术内容贡献句内容维度

（2）从位置分布来看，学术论文的贡献句立足于学术论文全文，可能出现在标题、摘要、引言、正文、结论等不同位置，即分布于全文的各个章节或段落。

（3）从研究价值与意义层面来看，贡献句是学术论文核心价值的文字体现，学术论文贡献句能让读者直接、准确地领略到作者研究贡献，具有传达创新观点、分享新成果、传播知识的功能。

3.2　数据选择与预处理

医学主题词表（Medical Subject Headings，MeSH）是一部由美国国立医学图书馆编制的规范化的可扩充的动态性叙词表，美国国立医学图书馆以其作为生物医学标引依据建立了国际上最权威的生物医学文献数据库——MEDLINE。PubMed 是互联网上使用最广泛的学术搜索引擎之一，提供生物医学论文和摘要数据检索服务，其数据来自 MEDLINE 数据库。MEDLINE 为其中收录的每篇文献提供了相对应的 MeSH 主题词，该主题词经过专家人工标注和标准规范化处理，是规范化的检索语言，能够集中体现文章的核心贡献内容，保证了主题词标注的准确性。MeSH 主题词表于 1989 年正式出版，为提高数据质量，笔者获取了 PubMed 数据库中 1989 年至 2015 年间所收录的论文全文和 MeSH tree 2015，构建了学术论文贡献句抽取数据集，包括论文的全文数据、题录信息和 Mesh 主题词，数据信息如图 2 所示：



id	PMCID	PMID	sentence	lemma_sentence	tag_sentence	tag	paper_i
284556	PMC3804280	24187659	In fact, Beauveria bassiana is one of the	in fact , beauveria bassiana b	in fact , #beauveria#TW bassiana be	TW TW VB	4299
284557	PMC3804280	24187659	Considering that there is a high genetic	consider that there be a high	consider that there be a #high#TW	TW TW TW	4299
284558	PMC3804280	24187659	This study was undertaken to assess the	this study be undertake to a	this #study#NN be undertake to as	NN TW TW	4299
284559	PMC3804280	24187659	The evaluation was performed in the sa	the evaluation be perform in	the #evaluation#NN be perform in	NN TW NN	4299
284560	PMC3804280	24187659	Biotic and abiotic factors have been rep	biotic and abiotic factor hav	biotic and abiotic factor have be #	VB TW TW	4299
284561	PMC3804280	24187659	Population levels of live adults of C. sor	population level of live adult	#population#TW level of #live#TW	TW TW TW	4299
284562	PMC3804280	24187659	However, mainly in Mutuípe, a higher a	however , mainly in mutuípe	however , mainly in mutuípe , a #hi	AD TW TW	4299
284563	PMC3804280	24187659	Higher mortality was found for chemica	higher mortality be find for	#higher#AD #mortality#TW be #fin	AD TW VB	4299
284564	PMC3804280	24187659	Since the plots presented similar condi	since the plot present simila	since the plot #present#VB similar	VB TW TW	4299
284565	PMC3804280	24187659	The mortality registered for the entom	the mortality register for the	the #mortality#TW register for the	TW TW VB	4299
284566	PMC3804280	24187659	Isolate CNPMF 218 was the most effect	isolate cnpmf 218 be the mo	isolate cnpmf 218 be the most #eff	TW TW TW	4299
284567	PMC3804280	24187659	Although those values are lower than t	although those value be low	although those value be #lower#TV	TW VB TW	4299
284568	PMC3804280	24187659	As mentioned before, studies conducte	as mention before , study co	as mention before , #study#NN cor	NN TW TW	4299
284569	PMC3804280	24187659	Thus, the biological control applied is a	thus , the biological control	thus , the #biological#TW #control	TW TW TW	4299
284570	PMC4022116	24877149	Based on recent advances in our under	base on recent advance in w	base on recent advance in #we#RF	RF TW RF \	4300
284571	PMC4022116	24877149	In this paper, we review recent progres	in this paper , we review rece	in this #paper#TW , #we#RF #revie	TW RF TW	4300
284572	PMC4022116	24877149	Nocturnal increase in AANAT enzymati	nocturnal increase in aa nat	nocturnal increase in aa nat enzyma	NN TW TW	4300
284573	PMC4022116	24877149	As mentioned above, several homeobo	as mention above , several h	as mention above , several #homeo	TW TW VB	4300
284574	PMC4022116	24877149	As reviewed below, a set of homeobox	as review below , a set of ho	as #review#TW below , a set of #ho	TW TW TW	4300
284575	PMC4022116	24877149	During development, Crx is expressed i	during development , crx be	during development , crx be expres	TW TW TW	4300
284576	PMC4022116	24877149	Around the same embryonic stage, Crx	around the same embryonic	around the same #embryonic#TW s	TW VB TW	4300
284577	PMC4022116	24877149	However, in a mouse with conditional C	however , in a mouse with c	however , in a #mouse#TW with co	TW TW TW	4300
284578	PMC4022116	24877149	Interestingly, several homeobox genes	interestingly , several homeo	interestingly , several #homeobox#	TW TW VB	4300
284579	PMC4022116	24877149	Investigations in a Crx-knockout mouse	investigation in a crx-knocko	#investigation#NN in a crx-knocko	NN TW VB	4300
284580	PMC4022116	24877149	In vitro studies have shown that the CR	in vitro study have show that	in vitro #study#NN have #show#VB	NN VB TW	4300
284581	PMC4022116	24877149	NRL and CRX have been shown to trans	nrl and crx have be show to	mammalian#TW #pineal gland#TW .	VB TW TW	4300
284582	PMC4022116	24877149	Thus, a similar cooperation between NF	thus , a similar cooperation b	thus , a similar cooperation betwee	NN TW TW	4300
284583	PMC4022116	24877149	In the adult rat pineal gland, many of th	in the adult rat pineal gland ,	in the #adult#TW #rat#TW #pineal	TW TW TW	4300
284584	PMC4022116	24877149	The daily expression profiles, existing d	the daily expression profile ,	the daily expression profile , exist #	NN TW TW	4300

图 2　学术论文贡献句抽取数据集（部分）

本研究的原始实验数据为英文格式的 nxml 文档，其中包含了无用的标签信息，需要进行文本清洗与预处理操作。通过编写相应规则将格式转换为纯文本格式，提取 nxml 格式中对应标题、正文、小节语义片段的

信息，删除多余的标签信息，利用停用词表去除噪音信息。构建 txt 格式的数据集后，为有效实现贡献句抽取，需要对数据进行预处理，包括对实验句子集中的每条句子进行分词，对文本序列进行词性标注，以及对文本进行词形还原，最终将处理后的数据储存在不同的文件中，供后续句子抽取使用。

3.3 基于 MeSH 主题词概念句抽取与筛选

单篇文章的 MeSH 主题词是领域专家对整篇论文研究工作的高度概括，涵盖了文献中已经涉及或可能相关的所有知识元，具有体现论文研究主题和贡献点的作用，主题词与句法特征的结合将为学术文本创新贡献句的挖掘提供重要线索。本研究采用的贡献句抽取方法的处理流程主要包括 MeSH 主题词提取、论文贡献句的引导词标注、基于规则的贡献句筛选 3 个步骤，如图 3 所示：

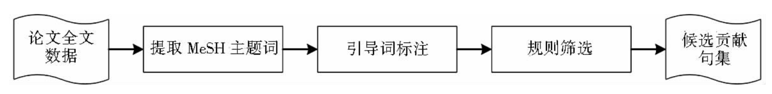


图 3 论文贡献句抽取流程

S. Mishra 等[29]以 MEDLINE 中收录的论文为研究对象，使用分配给文章的医学主题词（MeSH）标识，提出了科学论文中的时间新颖性、空间新颖性、组合时间新颖性、组合空间新颖性 4 种新颖性度量方法。笔者借鉴 S. Mishra 提出的度量方式，根据计算出的单篇文章中的 MeSH 词的新颖性得分对论文全文句子进行筛选，构建候选贡献句子集合。首先，从实验数据集中随机选择 4 300 篇文献，获取每篇文献的句子集和 MeSH 主题词集，共得到文献的句子集 658 760 句，平均每篇论文的句子数量为 153.2 句，单篇论文句子数最少为 92 句，最多为 281 句。其次，通过判断句子中是否含有该文章中新颖的主题词或主题词对，若包含则添加到候选贡献句集中，若不包含则将其舍去。最后，得到候选贡献句子集共 284 584 句，占总句数的 43.2%。本研究梳理了一套论文贡献句特征引导词表，共 6 类贡献句特征引导词，类型与示例如表 1 所示。6 类语言学特征引导词来自对医学研究领域的文献调研，并通过构建词表主体和分析医学领域贡献句词频，选取贡献句中的高频通用词加以补充。此外，领域主题词表由 MeSH tree 2015 解构得到，通过遍历树中所有根节点，即所有下位词，去掉重复部分，得到领域主题词表。

表 1 贡献句特征引导词表

类型	引导词示例
指代作者	I、we、our、the author 等
指代研究	this paper/article/publication/report/letter/context、this study/research、this contribution/method/result 等
特征动词	put foraward、find、reveal、illustrate、suggest、improve、design、develop、present、propose、shown 等
特征名词	insight、finding、analysis、investigation、solution、aim、objective、purpose、goal 等
特征形容词	novel、new、state of the art、better、stronger、unused 等
主题词	genome、bezoars、cardiology、myelophthisic 等

首先，根据 6 类特征引导词，对候选贡献句子集中的句子进行预处理，对词进行词性判断与词性还原。其次，采用最大双向匹配算法在不同类型的词表中对词进行查询匹配，对得到的结果进行比较，选择匹配后在词库中词数最多的方式为正确的分词方法，继而对实验句子进行标注。再次，将标注结果与词原型序列分离，得到句子标注序列。最后，编写正则表达式对句子标注序列进行匹配，抽取符合规则的句子标注序列及其对应的原文，从而得到候选贡献句集。

4 学术论文贡献句识别

4.1 实验设计与数据标注

本研究使用候选贡献句集合作为实验的标注对象，从 4 400 篇医学领域文献中随机选取 60 篇作为标注实验的样本，这 60 篇文献涵盖了多种不同的医学领域，在一定程度上保证了实验样本的广泛性。标注样本中共出现候选贡献句 2 936 条，其中包括正样本 2 034条，负样本 902 条。为了确保标注的客观性，笔者利用 Kappa 系数对标注结果进行一致性评估，选取了 3 位标注者共同标注的 15 篇文献（共包括 516 个句子）做交叉检验，计算得出 Kappa 一致性系数为 0.7。根据 S. Teufel[30]给出的一致性参考指标（K≥0.69，表示可靠），可以发现，本研究的标注结果达到了相对可靠的一致性水平。

论文贡献句抽取能够帮助读者了解作者的研究中取得了哪些成果。但是从文本内容角度来看，贡献句的判断往往依靠同行评议者的主观判断，受限于评议者的认知与经验。针对这一问题，本研究旨在有效利用论文全文和主题信息，为论文创新贡献句难于发现这一问题提供解决思路。首先，利用自然语言处理技术对全文本进行数据噪音去除、分词、去停用词等操作；其次，根据贡献句抽取规则并结合单篇文章的

Mesh 主题词,从全文本中抽取候选贡献句;再次,结合领域词表和特征词表,对抽取出来的候选贡献句进行筛选,形成候选贡献句集合;接着,采用半自动的方式对候选贡献句进行二分类标注,正样本为符合论文贡献句特征的句子,负样本为规则识别为贡献句而从实际上下文语境来看不符合实际贡献句的句子;最后,将标注的数据按照 6:1 的比例分配训练集和测试集,采用 Albert 模型训练贡献句识别模型,通过多次参数调整与模型优化,最终生成贡献句识别模型并实现对论文贡献句的识别。学术论文贡献句识别流程如图 4 所示:

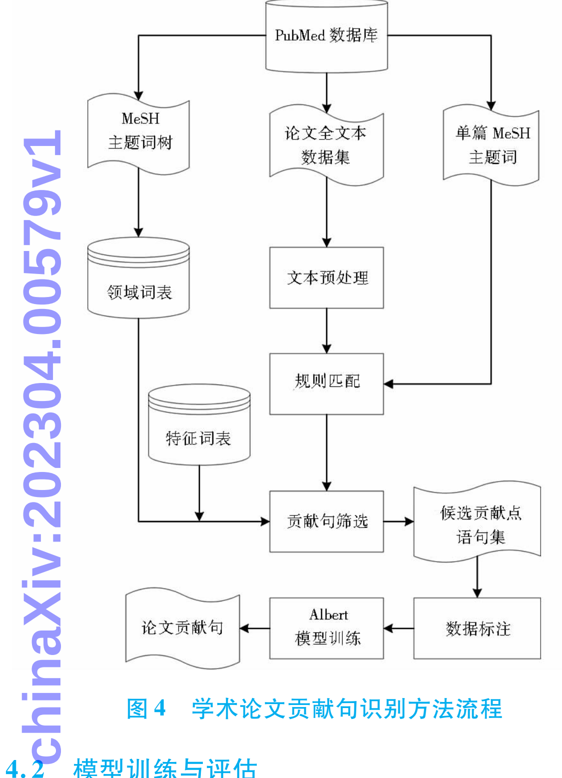


图 4 学术论文贡献句识别方法流程

4.2 模型训练与评估

完成数据的标注后,在训练集上用 Albert 模型进行模型训练,并在测试集上进行测试,本研究中所有的实验均在如表 2 所示的实验环境中完成。将 Albert 模型的 Batch size 设置为 32,设置最大句长为 256,不断调整实验的其他超参数,多次运行程序,记录在验证集上表现最优的组合。多次实验调参后将 globle_step 设置为 1 500,每隔 20 步保存一次模型,计算在测试集上的实验准确率。通过参数调整和多次实验测试,笔者发现无论是在训练集上还是测试集上,模型的准确率都在 90% 以上,达到了较高的水平。

此外,本研究还将 Albert 的实验结果与其他深度学习模型(FastText、TextCNN、TextRNN)和预训练模型(BERT 和 XLNet)进行对比实验,比较不同深度学习模型在贡献句分类上的具体表现。在参数设置过程中,

表 2 实验环境及配置

实验环境	环境配置
操作系统	Ubuntu16.04
GPU	NVIDIA Tesla T4
内存	32G
编程语言	Python3.6
深度学习框架	TensorFlow1.15
CUDA 版本	CUDA 11.1

本研究尽可能保证 6 种模型在参数设置上的一致性。在模型结构差异导致无法取得一致的情况下,取参数的最优设置。其他深度学习模型 FastText、TextCNN、TextRNN 采用的词向量维度为 128;预训练模型中,BERT 的注意力头数使用 8 头。为了避免结果的偶然性,多次运行程序,通过平均值进行比较,具体结果如表 3 所示:

表 3 不同模型在贡献句分类上的实验结果

分类模型	精确率/%	召回率/%	F1/%	准确率/%
Textcnn	79.54	80.45	79.99	79.71
Textrnn	81.31	81.47	81.39	80.90
Fasttext	80.93	82.97	81.94	81.86
Xlnet	89.64	90.59	90.11	89.73
BERT	88.77	89.82	89.29	89.49
Albert	91.97	91.52	91.74	91.60

4.3 实验结果分析

实验数据表明,针对学术论文中的贡献句识别问题,相对于传统深度学习方法,预训练模型在精确率、召回率、F1 值和准确率值上都取得了明显的优势,同时不同预训练模型间的差异较小,与在其他分类任务上的结果相似。其中,本研究训练的 Albert 模型的上述各项指标均取得最好的效果。

为进一步检验 Albert 模型的效果,假设 BERT、XLNet、Albert 这 3 种不同的深度学习模型性能相同,彼此之间不存在显著性差异,对模型的结果进行显著性假设检验。然而由于标注样本的数量有限,在使用交叉验证等实验估计方法时,如果每轮次选取的样本数较大,不同轮次的训练集会有一定程度的重叠,将导致假设成立的概率结果计算过高。如果每轮次选取的样本数较小,会使得结果的偶然性误差较大。为此,本文采用 5 * 2 交叉验证 t 检验法[31],将 3 种模型两两之间相互比较。5 * 2 交叉验证是做 5 次交叉 2 折交叉检验,在每次 2 折交叉验证之前,随机打乱数据的顺序,确保使得 5 次交叉验证中的数据划分不重复。第 i 次 2 折交叉验证将产生 2 对测试错误率。对 2 对测试错误率

分别求差,得到第一折上的差值 Δ_i^1 和第二折上的差值 Δ_i^2。为确保测试错误率的独立性,仅计算第 1 次 2 折交叉验证的两个结果平均值 $\mu = 0.5(\Delta_i^1 + \Delta_i^2)$。再对每次 2 折实验的结果计算其方差 $\sigma_i^2 = (\Delta_i^1 - 0.5\times(\Delta_i^1 + \Delta_i^2))^2 + (\Delta_i^2 - 0.5\times(\Delta_i^1 + \Delta_i^2))^2$,假设成立的概率为:

$$\tau_t = \frac{\mu}{\sqrt{0.2\sum_{i=1}^{5}\Sigma_i^2}} \quad 式(1)$$

实验结果表明 $\overline{\tau_t} = 2.561$,小于显著度 $\alpha = 0.05$ 时的临界值 2.570 6,大于显著度 $\alpha = 0.1$ 时的临界值 2.015 0,说明假设在显著度 $\alpha = 0.1$ 时不成立,即 3 种模型之间存在一定的显著性差异,其中平均错误率较小的 Albert 模型性能较优。

5 结语

学术论文贡献点的自动识别是实现论文智能评价的重要环节,对科技评价工作的开展具有导向和推动作用。本研究针对目前贡献句抽取研究存在的不足,以 MEDLINE 数据库中的期刊文献、MeSH 主题词为基础,引入深度学习、文本分析等领域的理论与技术,从句子层面对论文贡献内容进行挖掘和分析,提出了学术论文贡献句识别方法。本研究旨在从学术论文全文本中抽取完整意义的贡献句,揭示论文的贡献点,为实现更加语义化、智能化的学术论文创新性评价奠定基础。通过实验验证和对比分析,证明了本研究采用的 Albert 模型的合理性以及该模型在处理贡献句分类问题上的优越性。

本研究的意义在于通过上述方法可以准确地自动识别文章中的贡献句子,突出论文的创新性工作。一方面,可以降低同行评议中的审稿压力,在创新知识传播、研究方向梳理等方面具有较高的应用价值;另一方面,为从论文句子内容层面评价论文创新性做出了尝试和基础铺垫,为构建学术论文创新点识别和创新性评价研究奠定了基础。

参考文献:

[1] 新华网. 习近平:在科学家座谈会上的讲话[EB/OL]. [2021－05－07]. http://www.xinhuanet.com/2020-09/11/c_1126483997.htm.
[2] 国家标准化管理委员会. 科学技术报告、学位论文和学术论文的编写格式:GB 7713－87 [S]. 北京:中国标准出版社,1987.
[3] 李如森,彭彩红,赵福荣. 科技论文创新性判断方法[J]. 鞍山钢铁学院学报,2001(3):234－236.
[4] 温有奎,吴广印. 碎片化科研创新点动态挖掘研究[J]. 数字图书馆论坛,2014(7):25－32.
[5] 张帆,乐小虬. 面向领域科技文献的句子级创新点抽取研究[J]. 现代图书情报技术,2014(9):15－21.
[6] 索传军,于果鑫. 学术论文研究亮点的语言学特征与分布规律研究[J]. 图书情报工作,2020,64(9):104－113.
[7] 章成志,李铮. 基于学术论文全文的创新研究评价句抽取研究[J]. 数据分析与知识发现,2019,3(10):12－19.
[8] 曹树金,闫欣阳,张倩,等. 中外情报学论文创新性特征研究[J]. 图书情报工作, 2020,64(1):80－92.
[9] 温浩. 科技文摘创新点语义识别与分类方法研究[J]. 情报学报,2019,38(3):249－256.
[10] 周海晨,郑德俊,郦天宇. 学术全文本的学术创新贡献识别探索[J]. 情报学报,2020,39(8):845－851
[11] CHEN L L, FANG H. An automatic method for extracting innovative ideas based on the Scopus® database[J]. Knowledge organization, 2019, 46(3): 171－186.
[12] ALLAN J, WADE C, BOLIVAR A. Retrieval and novelty detection at the sentence level[C]//Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. Toronto:ACM,2003: 314－321.
[13] TEUFEL S, MOENS M. Summarizing scientific articles: experiments with relevance and rhetorical status[J]. Computational linguistics, 2002, 28(4): 409－445.
[14] HEFFERNAN K, TEUFEL S. Identifying problems and solutions in scientific text[J]. Scientometrics, 2018, 116(2): 1367－1382.
[15] 冷伏海,白如江,祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作,2013,57(11):112－119.
[16] 毛琛瑜,乐小虬. 领域内中文科技文献中新发现语言描述特征分析[J]. 现代图书情报技术,2016(5):47－55.
[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2021－05－07]. https://arxiv.org/pdf/1301.3781v3.pdf.
[18] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]// Proceedings of the 2014 conference on empirical methods in natural language processing. Doha: Association for Computational Linguistics, 2014: 1532－1543.
[19] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, Volume 1 (long papers). New Orleans: Association for Computational Linguistics, 2018: 2227－2237.
[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers). Minneapolis: As-

sociation for Computational Linguistics, 2019:4171 – 4186.
[21] 鲁威. 基于多因素特征的文本分类的研究[D]. 成都:电子科技大学,2019.
[22] 顾亦然,霍建霖,杨海根,等. 基于 BERT 的电机领域中文命名实体识别方法[EB/OL]. [2021 – 05 – 07]. https://doi. org/10.19678/j. issn. 1000-3428. 0058838.
[23] 廖胜兰,吉建民,俞畅,等. 基于 BERT 模型与知识蒸馏的意图分类方法[EB/OL]. [2021 – 05 – 07]. https://doi. org/10.19678/j. issn. 1000-3428. 0057416.
[24] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[EB/OL]. [2021 – 05 – 07]. https://openreview. net/pdf? id = H1eA7AEtvS.
[25] DENHOLM C J, PHILPOTT C. Making the implicit explicit: creating performance expectations for the dissertation[J]. Quality assurance in education, 2009, 17(2):204 – 206.
[26] DAHL T. Contributing to the academic conversation: a study of new knowledge claims in economics and linguistics[J]. Journal of pragmatics, 2008, 40(7): 1184 – 1201.
[27] 李瑛,周立. 科技期刊论文创新点合理呈现的价值及理想模式[J]. 中国科技期刊研究,2018,29(10):993 – 999.
[28] 李贺,杜杏叶. 基于知识元的学术论文内容创新性智能化评价研究[J]. 图书情报工作, 2020,64(1):93 – 104.
[29] MISHRA S, TORVIK V I. Quantifying conceptual novelty in the biomedical literature [EB/OL]. [2021 – 05 – 07]. http://www. dlib. org/dlib/september16/mishra/09mishra. html.
[30] TEUFEL S, SIDDHARTHAN A, TIDHAR D. An annotation scheme for citation function[C]// Proceedings of the 7th SIGdial workshop on discourse and dialogue. New York: ACM, 2006: 80 – 87.
[31] DIETTERICH T G. Approximate statistical tests for comparing supervised classification learning algorithms[J]. Neural computation, 1998, 10(7): 1895 – 1923.

作者贡献说明:
罗卓然:确定论文思路,设计实验方案,撰写与修改论文;
蔡乐:数据标注与实验分析;
钱佳佳:数据标注与实验分析;
陆伟:提出研究问题,修改论文。

Research on the Recognition of Innovative Contribution Sentences of Academic Papers

Luo Zhuoran Cai Le Qian Jiajia Lu Wei
School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Contribution sentences of academic papers are elements to reflect the novelty and academic value of papers. This study takes the full text of academic papers and MeSH terms as data sources and uses natural language processing and deep learning techniques to achieve academic paper contribution sentence recognition. This study lays the foundation for fine-grained mining of innovative contents of academic texts, which is important for realizing the evaluation of academic papers based on cognitive computing. [**Method/process**] Firstly, the full-text PubMed papers were used as the data source for element analysis and feature extraction of the contributed sentences. Secondly, a semi-automatic approach was used to fulfill the data annotation. Finally, the automatic recognition of contributed sentences was realized based on Albert deep learning model. [**Result/conclusion**] The plausibility of the experimentally labeled training data is proved by the data consistency test, and the experimental results show that the automatic recognition model trained in this paper can identify the contribution sentences in academic papers more effectively compared with other deep learning models.

Keywords: contribution sentences academic papers novelty Albert